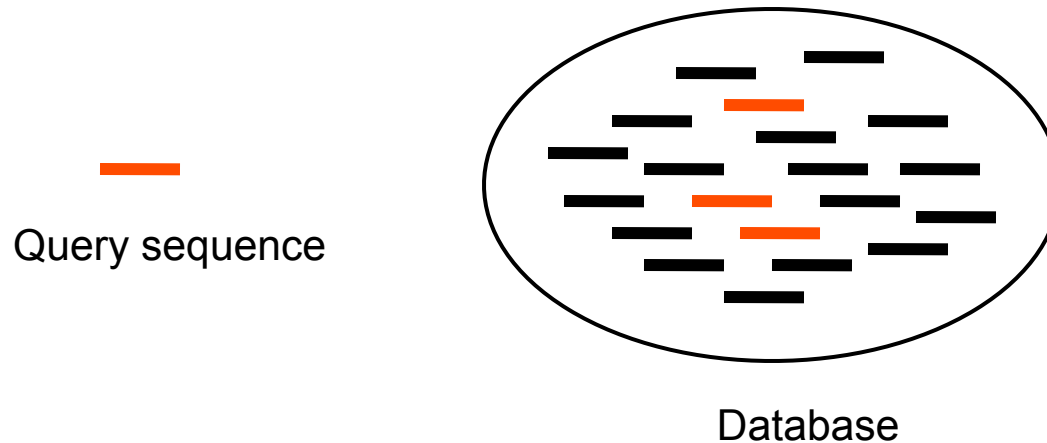

BLAST

Anders Gorm Pedersen
&
Rasmus Wernersson

Database searching

**Using pairwise alignments to search
databases for similar sequences**



Database searching

Most common use of pairwise sequence alignments is to search databases for related sequences. For instance: find probable function of newly isolated protein by identifying similar proteins with known function.

Most often, ***local*** alignment (“Smith-Waterman”) is used for database searching: you are interested in finding out if ANY domain in your protein looks like something that is known.

Often, full Smith-Waterman is too time-consuming for searching large databases, so heuristic methods are used (fasta, BLAST).

Database searching: heuristic search algorithms

FASTA (Pearson 1995)

Uses heuristics to avoid
calculating the full dynamic
programming matrix

Speed up searches by an order of
magnitude compared to full
Smith-Waterman

BLAST (Altschul 1990, 1997)

Uses rapid word lookup methods
to completely skip most of the
database entries

Extremely fast

One order of magnitude
faster than FASTA

Two orders of magnitude
faster than Smith-
Waterman

Almost as sensitive as FASTA

BLAST flavors

BLASTN

Nucleotide query sequence

Nucleotide database

BLASTP

Protein query sequence

Protein database

BLASTX

Nucleotide query sequence

Protein database

Compares all six reading frames
with the database

TBLASTN

Protein query sequence

Nucleotide database

"On the fly" six frame translation of
database

TBLASTX

Nucleotide query sequence

Nucleotide database

Compares all reading frames of
query with all reading frames of
the database

Searching on the web: BLAST at NCBI

Very fast computer dedicated to running BLAST searches

Many databases that are always up to date (e.g. NR and Human Genome)

Nice simple web interface

The screenshot displays the NCBI BLAST web interface in a browser window. The address bar shows the URL: <http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&...>. The page title is "Protein BLAST: search protein databases using a protein query". The interface includes a navigation bar with links: Home, Recent Results, Saved Strategies, and Help. A "My NCBI" link is also present. The main content area is titled "Enter Query Sequence" and contains several input fields: "Enter accession number, gi, or FASTA sequence" (with a "Clear" button), "Query subrange" (with "From" and "To" fields), "Or, upload file" (with a "Choose File" button and "no file selected" text), and "Job Title" (with a text input field). Below these fields is a "Choose Search Set" section with "Database" (set to "Non-redundant protein sequences (nr)"), "Organism" (with a text input field and a note: "Enter organism name or id--completions will be suggested. Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown."), and "Entrez Query" (with a text input field and a note: "Enter an Entrez query to limit search"). The "Program Selection" section shows "Algorithm" with radio buttons for "blastp (protein-protein BLAST)" (selected), "PSI-BLAST (Position-Specific Iterated BLAST)", and "PHI-BLAST (Pattern Hit Initiated BLAST)". A "BLAST" button is prominently displayed. Below it, a checkbox for "Show results in a new window" is visible. The footer contains links for "Copyright", "Disclaimer", "Privacy", "Accessibility", "Contact", and "Send feedback on new interface", along with "NCBI | NLM | NIH | DHHS".

When is a database hit meaningful?

- **Problem:**

- Even unrelated sequences can be aligned (yielding a low score)
- How do we know if a database hit is meaningful?
- When is an alignment score sufficiently high?

- **Solution:**

- Determine the range of alignment scores you would expect to get for random reasons (i.e., when aligning unrelated sequences).
- Compare actual scores to the distribution of random scores.
- Is the real score much higher than you'd expect by chance?

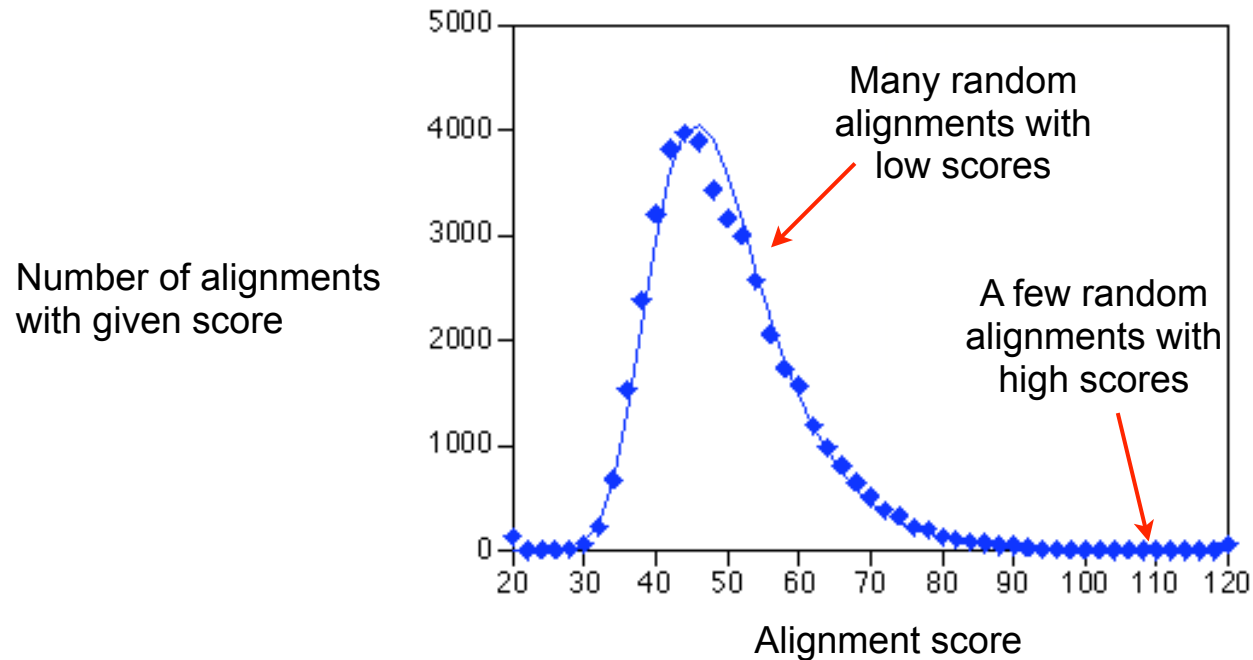
Distribution of random alignment scores

- Software simulation

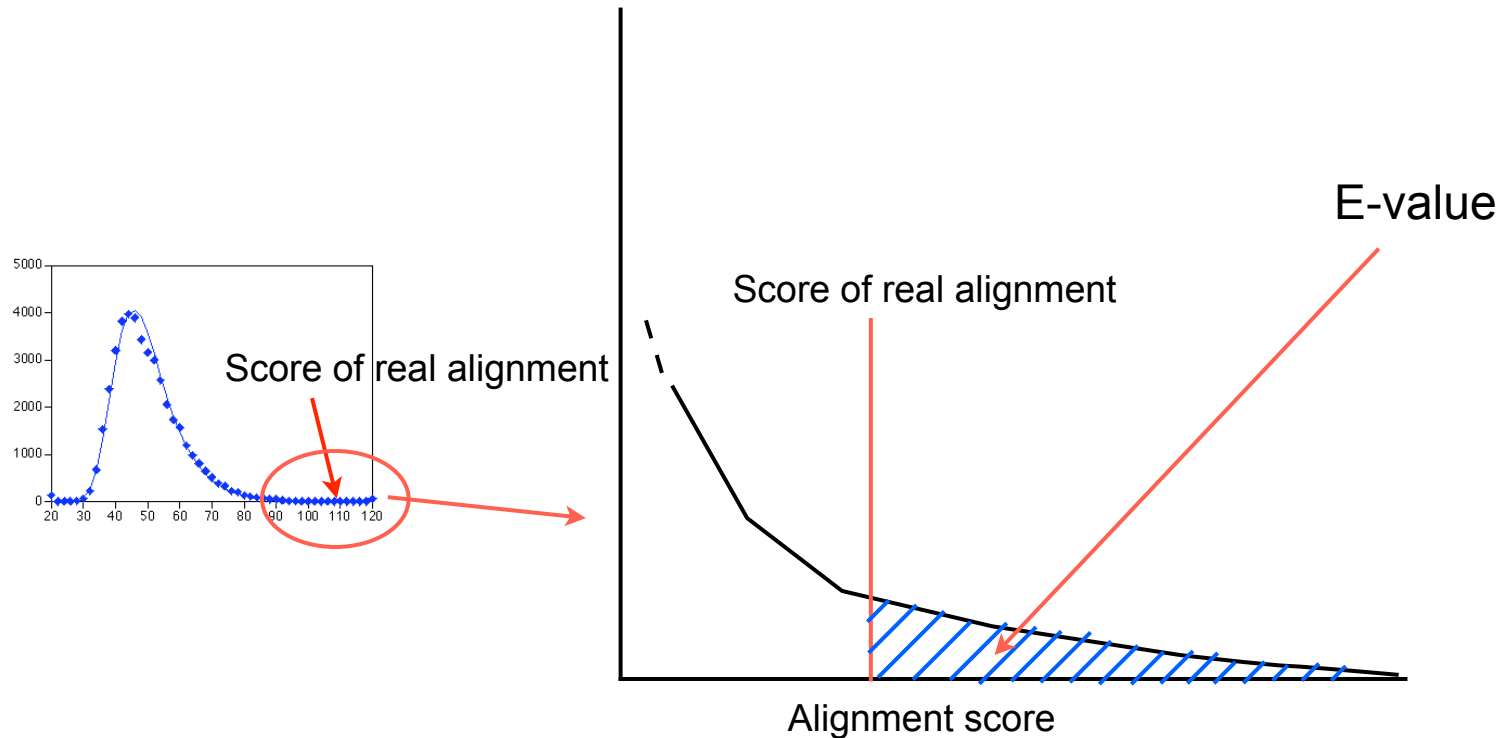
Significance of alignment score expressed as E-value

Pairwise alignment of unrelated sequences results in scores following an extreme value distribution

The exact shape and location of the distribution depends on the length and composition of the sequences



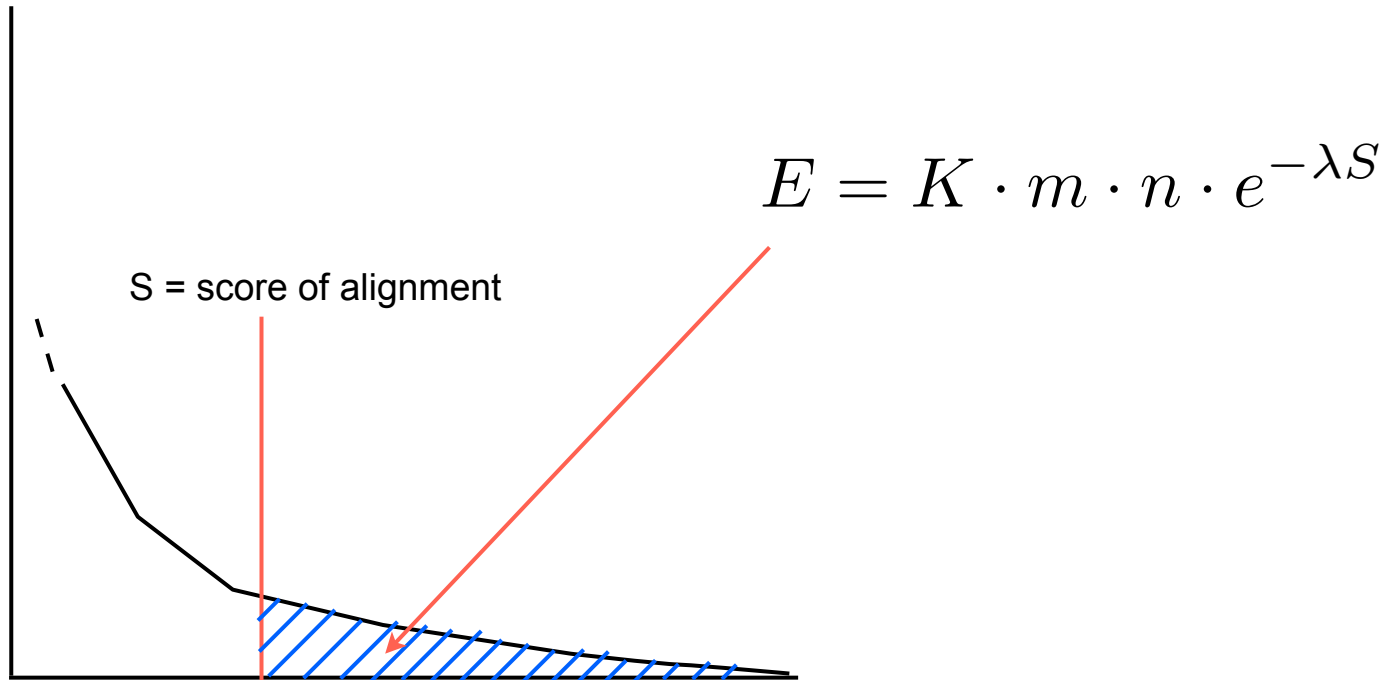
Significance of alignment score expressed as E-value



E-value: the number of random hits with score \geq real score

Want E-values well below 1 (the lower the better)

Significance of alignment score expressed as E-value



m : length of query sequence

n : combined length of all database sequences

λ : the scaling factor we also saw when computing BLOSUM62

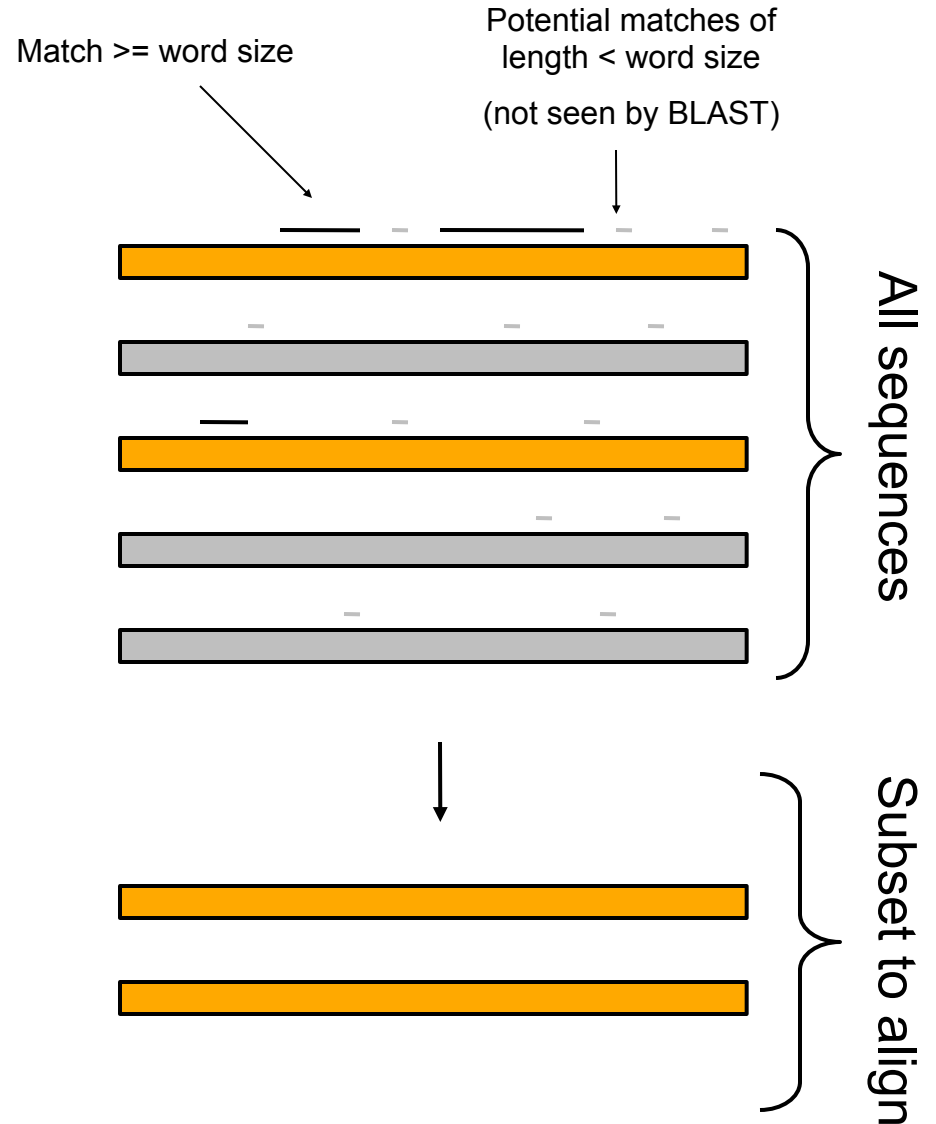
K : a constant whose value depends on the nature of the sequences - it can be determined empirically by curve fitting

BLAST heuristics

- BLAST speeds up the search $>100x$ by pre-screening the database sequences and only aligning “*promising*” sequences.
- Promising sequences: database sequences that have sub-strings (“words”) which also occur in the query sequence (found rapidly using a so-called “**suffix-tree**”)
- **BLASTN** and **BLASTP** use different criteria for overlap required for a sequence to be deemed promising

BLASTN

- Heuristics:
 - Perfectly matching “word” of size ≥ 11 amino acids
- DNA alignment matrix:
 - Match: **1**
 - Mismatch: **-3**



BLASTP

- Heuristics:
 - 2 x “near match” within a window.
 - These matches must be on the same alignment “diagonal” (i.e., if the words are N residues apart in the query, they also have to be N residues apart in the target sequence).
 - Default word length: 3 aa
 - Default window length: 40 aa
- Alignment matrix:
 - Default: BLOSUM 62

